BOLT BERANEK AND NEWMAN INC

CONSULTING • DEVELOPMENT • RESEARC

BBN Report No. 2154 10 June 1971

STRATEGIES FOR RECOGNITION OF SPOKEN SENTENCES FROM VISUAL EXAMINATION OF SPECTROGRAMS

by

Dennis H. Klatt and Kenneth N. Stevens

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

This research was supported by the Advanced Research Projects Agency under ARPA Order No. 1967; Contract No. DAHC-71-C-0088.

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce for sale to the general public.

37

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Bolt Beranek and Newman Inc. 50 Moulton Street Cambridge, Massachusetts 02138	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE
STRATEGIES FOR RECOGNITION OF SPOKEN SENTENCES FROM VISUAL EXAMINATION OF SPECTROGRAMS

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Scientific

5. AUTHOR(S) *(First name, middle initial, last name)*
Dennis H. Klatt
Kenneth N. Stevens

6. REPORT DATE 10 June 1971	7a. TOTAL NO. OF PAGES 33	7b. NO. OF REFS 25

8a. CONTRACT OR GRANT NO. DAHC15 71 C 0088 b. PROJECT NO ARPA ON 1967 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) BBN Report No. 2154 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce for sale to the general public.

11. SUPPLEMENTARY NOTES This research was sponsored by the Advanced Research Projects Agency under ARPA Order No. 1967.	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

The aim of this study is to gain insight into the strategies that might be necessary in a device for the automatic recognition of spoken sentences, through an experiment in which speech recognition is attempted by visual recognition of spectrograms by experienced experimenters. Spectrograms of a set of ten sentences, constructed from a vocabulary of 200 words, were prepared and the experimenters (the authors) attempted two tasks from visual examination of the spectrograms: (1) phonetic transcription of the sentences in terms of phonetic symbols or in terms of a partial feature specification; and (2) recognition of each sentence as a whole, using any available information, including the lexicon.

In the phonetic transcription task, 56 percent of the segments were recognized correctly, and the feature specification was partially correct in an additional 27 percent of the segment. In the sentence recognition task the experimenters missed 27 words out of a possible 156, but most of these were simple function words. The sentence-recognition strategies used by the experimenters consisted of three steps: 1. Identification of clear and well defined phonetic segments and features; 2. Hypothesis of remaining features by reference to the lexicon and syntactic and semantic constraints; and 3. Reexamination of the acoustic data to see if they are consistent with the hypothesis.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
speech recognition phonetic transcription speech analysis speech spectrograms						

STRATEGIES FOR RECOGNITION OF SPOKEN SENTENCES FROM VISUAL EXAMINATION OF SPECTROGRAMS*

Dennis H. Klatt and Kenneth N. Stevens**

*A condensed version of this paper was presented at the 81st meeting of the Acoustical Society of America in Washington, D.C., on April 20-23, 1971.

**The authors are also at the Massachusetts Institute of Technology.

ABSTRACT

The aim of this study is to gain insight into the strategies that might be necessary in a device for the automatic recognition of spoken sentences, through an experiment in which speech recognition is attempted by visual recognition of spectrograms by experienced experimenters. Spectrograms of a set of ten sentences, constructed from a vocabulary of 200 words, were prepared and the experimenters (the authors) attempted two tasks from visual examination of the spectrograms: (1) phonetic transcription of the sentences in terms of phonetic symbols or in terms of a partial feature specification; and (2) recognition of each sentence as a whole, using any available information, including the lexicon.

In the phonetic transcription task, 56 percent of the segments were recognized correctly, and the feature specification was correct but incomplete in an additional 27 percent of the segments. In the sentence recognition task, the experimenters missed 27 words (out of a possible 156), but most of these were simple function words. The sentence-recognition strategies used by the experimenters consisted of three steps: (1) Identification of clear and well defined phonetic segments and features; (2) Hypothesis of remaining features by reference to the lexicon and syntactic and semantic constraints; and (3) Reexamination of the acoustic data to see if they are consistent with the hypothesis.

It is suggested that similar procedures will be necessary in an automatic speech recognition task, and it is felt that this task is sufficiently complex that only simple and highly constrained sentences will be capable of recognition in the near future.

TABLE OF CONTENTS

BLANK PAGE

1.0 INTRODUCTION

During the past two decades there has been a continuing interest in the development of machines for the recognition of speech. The literature has been reviewed by Lindgren (1965) and, more recently, by Hyde (1968). This work has been concerned primarily with the recognition of words spoken in isolation, but recently there has been increasing emphasis on the recognition of sentence material (Sakai and Doshita, 1962; Martin, *et al.*, 1966; Reddy, 1967; Vicens, 1969; Tappert, *et al.*, 1970; Newell, (in press).

There are remarkable differences between an utterance of continuous speech and the same words spoken in isolation:

(1) Word boundaries are not clearly marked.

(2) Co-articulation occurs between words.

(3) Stress and syntactic information are encoded by modifications in segmental durations, fundamental frequency changes, pauses, and the reduction of vowels.

(4) The acoustic attributes that signal the feature values of many segments are changed, or some features may actually be deleted according to specific rules of English phonology. It is reasonable to suppose that a speaker is free to delete certain acoustic cues from sentence material since he is aware that the listener has available to him contextual information that enables him to supply the missing cues through some kind of internal calculation.

Some of the effects of putting words together into meaningful sentences are illustrated in Figs. 1 and 2. Figure 1 compares a broadband spectrogram* (Presti, 1966; Koenig, Dunn and Lacey, 1946) of an utterance with the same words when spoken in isolation. The sentence is "If the cube is not blue, pick it up." It is instructive to compare each isolated word with the portion of the sentence corresponding to the same word. The changes that one sees are sufficiently dramatic that, in general, one cannot hope to achieve sentence recognition by matching a set of stored acoustic patterns corresponding to isolated words against a comparable acoustic representation of the unknown utterance.

The assertion that even the most sophisticated acoustic pattern recognition schemes are insufficient for continuous speech recognition is reinforced by the example shown in Fig. 2. A spectrogram of the word "after" is compared with spectrograms of the same word embedded in several different sentences. The sentences have been chosen to illustrate co-articulation of the word "after" with adjacent vowels, sonorants, nasals, plosives and fricatives. As can be seen from these examples, vowels and other sonorants tend to produce greater acoustic changes than plosives and fricatives.

Since the recognition of sentences will clearly require strategies that are considerably more complex than those used in isolated-word recognition, it seems prudent to study the performance of a human observer when he is faced with the task of understanding sentences in the form of visual patterns that involve a transformation of the speech input similar to that

*A spectrogram plots time on the horizontal axis and frequency on the vertical axis. The blackness of any point is monotonically related to the energy contained in the previous 3-5 msec of the waveform obtained from the output of a 300 Hz bandpass filter centered at that point.

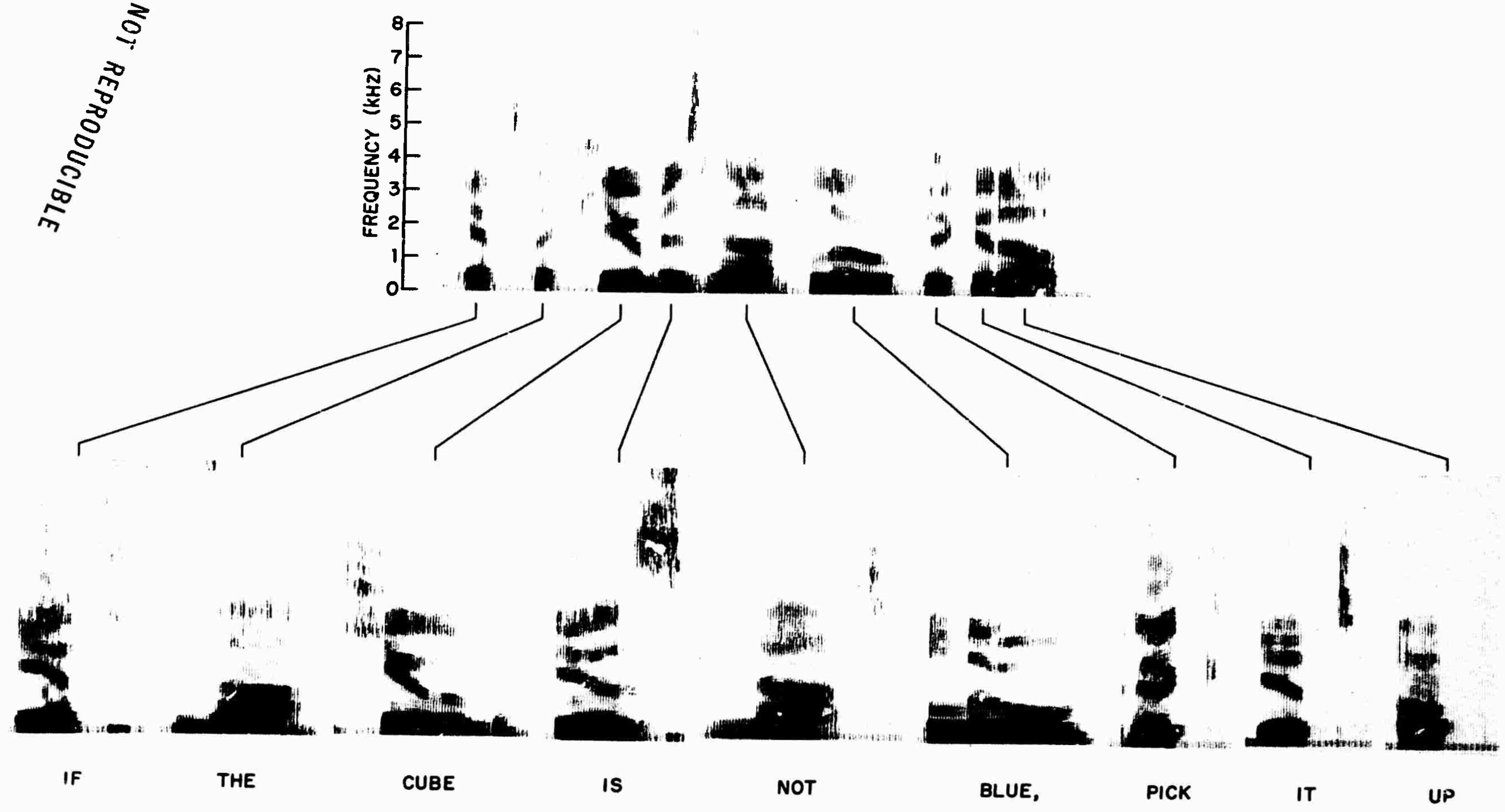
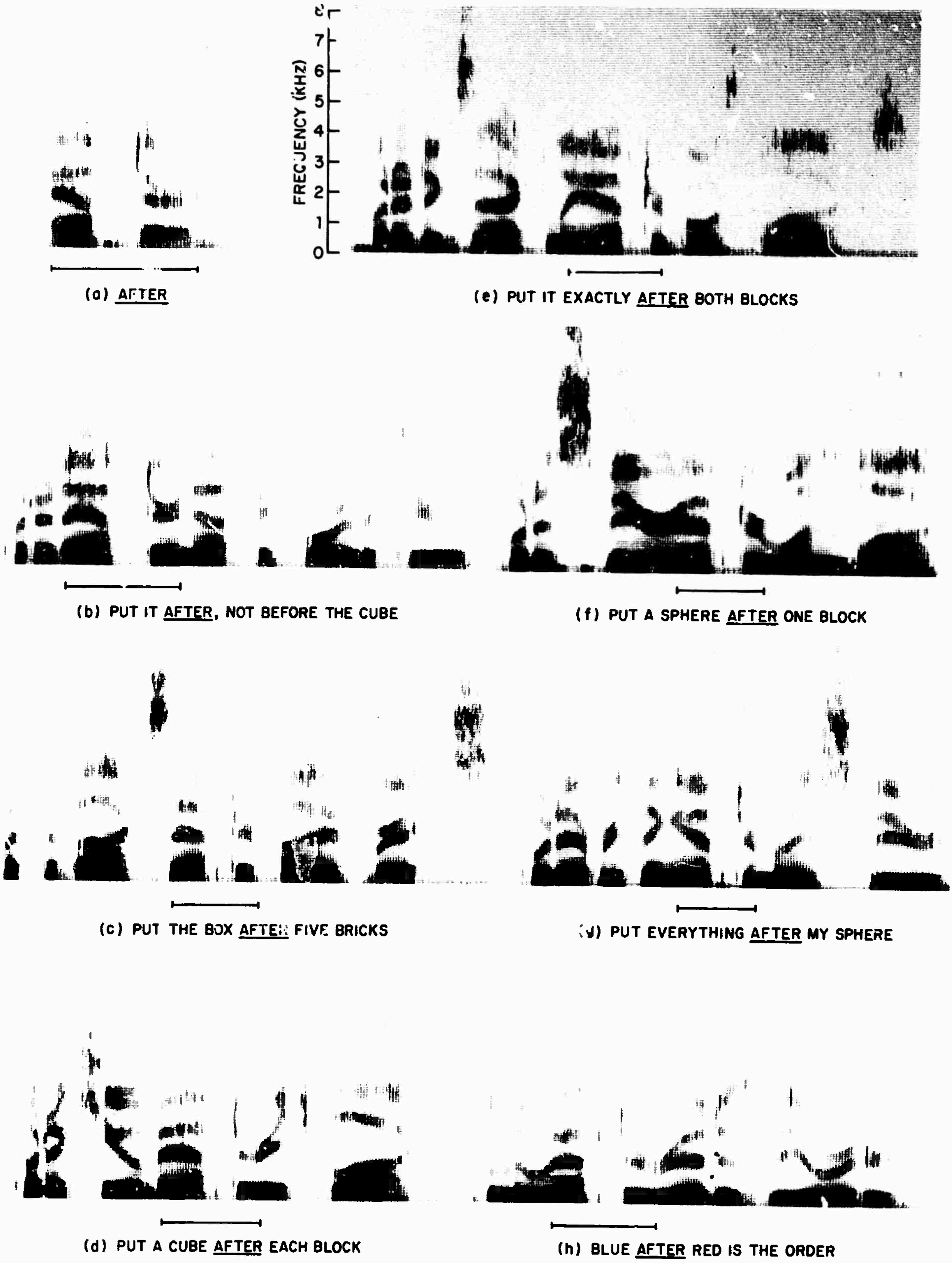


FIG.1 A BROADBAND SPECTROGRAM OF THE SENTENCE "IF THE CUBE IS NOT BLUE, PICK IT UP" IS SHOWN IN THE TOP HALF OF THE FIGURE. SPECTROGRAMS OF THE SAME WORDS SPOKEN IN ISOLATION APPEAR BELOW.



(a) AFTER

(e) PUT IT EXACTLY AFTER BOTH BLOCKS

(b) PUT IT AFTER, NOT BEFORE THE CUBE

(f) PUT A SPHERE AFTER ONE BLOCK

(c) PUT THE BOX AFTER FIVE BRICKS

(g) PUT EVERYTHING AFTER MY SPHERE

(d) PUT A CUBE AFTER EACH BLOCK

(h) BLUE AFTER RED IS THE ORDER

FIG.2 A SPECTROGRAM OF THE WORD "AFTER", SPOKEN IN ISOLATION, IS COMPARED WITH SEVERAL SENTENCES CONTAINING THE SAME WORD

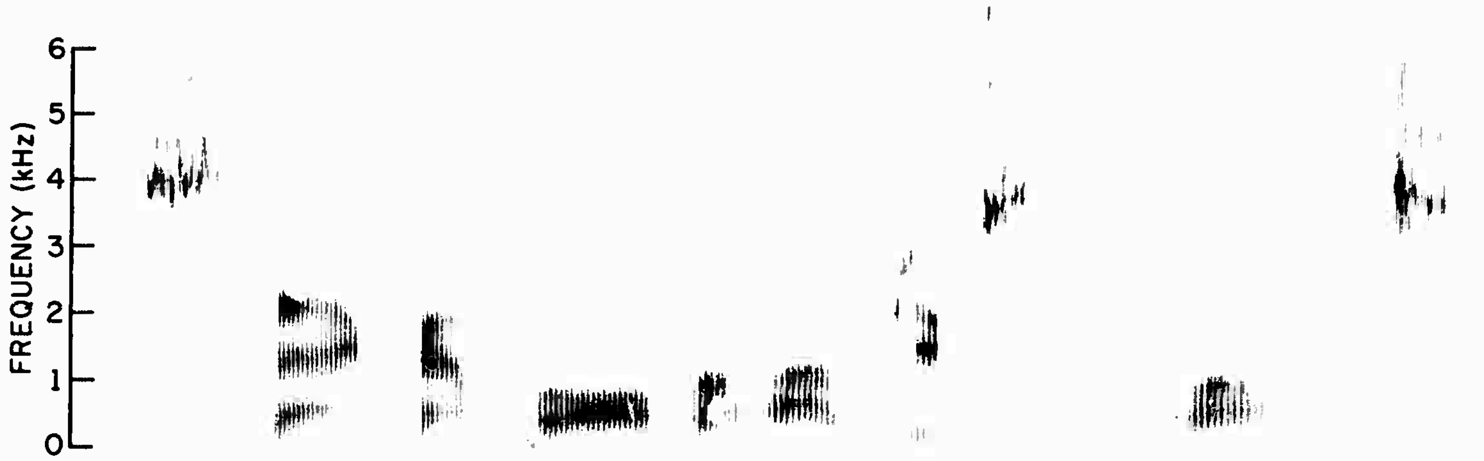
presented to a potential automatic recognition device. The two tasks that we have chosen are phonetic transcription and complete recognition of a set of spoken English sentences from visual examination of spectrograms. The problem of automatic extraction of acoustic properties such as formant frequencies, fundamental frequency, and rapid spectral changes is bypassed by identifying these properties visually from a broadband spectrographic representation.

If it can be demonstrated that subjects can learn to read spectrograms, then it can be concluded that the data available on a spectrogram at least provide a sufficient input for an automatic sentence recognition device, although there may, of course, be other transformations of the input data that would help to simplify the task. Apart from this observation, however, it would be a mistake to interpret success on the spectrogram-reading task as a positive indication of the potential success of automatic procedures. A human observer can bring to the spectrogram-reading task the enormously complex information-processing abilities at the semantic and syntactic levels that he uses in dealing with language—the kind of knowledge that he utilizes when speaking and listening, when reading, or when translating text from one language into another. In addition, the sophisticated observer can, through covert or overt internal generation of sentence material that he hypothesizes to be represented in a spectogram, verify phonetic facts which are otherwise not readily accessible to him. Thus, the spectrogram-reading exercise reported here should be interpreted not as a means for assessing the potential success of future speech-recognition devices but as a vehicle for gaining insight into the strategies that might be reasonable to follow in such devices.

Experiments on visual recognition of spectrograms have been performed for words spoken in isolation (Potter, Kopp and Green, 1947). Working with spectrograms and a real-time display, the authors were able to train several observers to recognize a lexicon of up to 200 common words from the visual representation. All words were spoken very distinctly, and silent pauses appeared between the words of a sentence or phrase. It is not known whether observers developed an ability to analyze an unfamiliar word by phonetic decomposition or to deal with normal continuous speech in real time, but recent research on speech analyzing aids for the deaf casts doubt on these possibilities (Goldberg, 1970; Liberman, *et al.*, 1968).

Spectrograms of isolated words have been used in informal visual recognition experiments in the past (Stevens, 1969). Broadband spectrograms were made of words selected from a 64-word lexicon spoken by several talkers (Gold, 1966). After a short period of instruction, students working in small groups were able to identify words correctly 85-100 percent of the time when provided with a list of the lexicon. Spectrograms have also been examined in attempts to recognize speaker identity. (Kersta, 1962; Tosi, *et al.*, 1971).

NOT REPRODUCIBLE



TRANSCRIPTION OF DHK:	s	t	æ	k	ʌ	ƀ	ɔ	l	b	ə	h	{ʌ a}		ǰ	ɪ	k s	{n m}	ə	b	a	{p t k}	s
"CORRECT" TRANSCRIPTION:	s	t	æ	k	ʌ	p	ɔ	l	b	l	u	a	b	ǰ	ɪ	k s	n	ə	b	a	k	s

WORD GUESSES OF DHK:	STACK	UP	ALL	THE	OBJECTS	IN	A	BOX
CORRECT WORD SEQUENCE:	STACK	UP	ALL	BLUE	OBJECTS	IN	A	BOX

FIG.3 A SPECTROGRAM OF ONE OF THE UNKNOWN SENTENCES IS SHOWN AND THE PHONETIC TRANSCRIPTION OF DHK IS COMPARED WITH THE "CORRECT" TRANSCRIPTION. THE ACTUAL WORD SEQUENCE AND D"K's GUESS AT THE WORD SEQUENCE ARE ALSO PRESENTED

2.0 PHONETIC TRANSCRIPTION TASK

The materials used in the first experiment were broadband spectorgrams of ten sentences spoken by a single male talker at a conversational rate. A sample spectrogram is shown in Fig. 3. One of the authors (DHK) attempted to make a transcription of the utterances in terms of phonetic segments, or, if some features were ambiguous, in terms of a partial feature specification of the segments.* During this phase of the experiment, neither the lexicon nor the semantic context of the sentences was known to the experimenter, and he tried to avoid making hypotheses about these matters. No spectrograms of this speaker had been observed previously.

An example of the phonetic transcription produced for the sentence "Stack up all blue objects in a box" is shown in Fig. 3 immediately below the spectrogram. Sets of phonetic symbols appearing within brackets are used as an abbreviation for the fact that one or more features could not be identified at all from the spectrogram. Thus the notation $\left\{\begin{smallmatrix}m\\n\end{smallmatrix}\right\}$ means that the segment identity was ambiguous as far as the labial or coronal features are concerned.

A so-called correct phonetic transcription appears below the transcription of the experimenter in Fig. 3. Determination of the correct transcription involves many rather arbitrary decisions, but we believe that these decisions have relatively little effect on the results to be reported.

*A discussion of what is meant by a phonetic segment and a feature is given in Section 3.2.3.

A summary of the transcription results is shown in Table 1. Approximately 200 phonetic segments should have been detected and transcribed. As the results indicate, one or more feature entries were left unspecified in over one quarter of the phonetic segments. The errors that were made were of three types. A segment was identified incorrectly on at least one feature dimension 11 percent of the time. A segment was not detected 5 percent of the time, and a segment was seen and transcribed when, in fact, none was present one percent of the time.

Correct and complete feature specification	=	56%	83%
Correct but partial feature specification	=	27%	
Incorrect at least one feature	=	11%	
Segment not even detected	=	5%	
Segment added	=	1%	

TABLE 1. An Error Analysis for the Phonetic Transcription of Ten Sentences

3.0 SENTENCE RECOGNITION TASK

The same ten spectrograms were used in the second experiment, whose goal was to recognize the correct word sequence. Each sentence contained an average of 8 words selected from a 200-word lexicon, plus plurals. The lexicon is listed in Table 2. It was obtained from Terry Winograd (1971) and is capable of describing and manipulating a scene containing objects in various relations to one another. The authors worked independently with the aid of the lexicon and the knowledge that the sentences were meaningful and well formed. The ten sentences were processed in about 3 hours.

3.1 Summary of Results

The results are shown in Table 3. The ten sentences are listed and words not identified correctly are underlined. DHK missed 16 words and KNS missed 11 words, but only 3 words were missed by both experimenters.

An analysis of the errors is shown in Table 4. Eight of the word errors were relatively minor "a - the" confusions. If these errors are disregarded, there remain 10 out of 20 sentences with at least one word incorrectly transcribed. However, if we had compared transcriptions before grading the results, we would probably have done much better. The data in Table 3 suggest that 9 out of the 10 sentences might have been recognized from our pooled knowledge.

A
ABOVE
AFTER
ALL
AN
AND
ANY
ANYTHING
ARE
AS
AS
ASK
AT
AWAY
BALL
BACK
BE
BEFORE
BEGIN
BEGAN
BEHIND
BELOW
BENEATH
BESIDE
BIG
BLACK
BLOCK
BLUE
BOTH
BOX
BRICK
BUILD
BOTH
BUT
BY
CALL
CAN
CHOOSE
CLEAN
CLEAR
COLOR
CONSTRUCT
CONTAIN
CORNER
CUBE
DID
DO
DOES
DOWN
DROP
EACH
EITHER
EVERY
EVERYTHING
EXACTLY
FEW
FEWER
FIND
FINISH
FIVE
FOUR
FRIEND
FROM
FRONT
GAVE
GIVE
GO
GOING
GRAB
GRASP
GREATER
HAD
HAND
HANDLE
HAS
HAVE
HE
HER
HIGH
HIM
HIS
HOLD
HOW
I
IF
IN
INSIDE
INTO
IS
IT
ITS
KNOW
LARGE
LEAST
LEFT
LESS
LIKE
LITTLE
LONG
MAKE
MANY
ME
MORE
MOST
MOVE
MY
NAME
NARROW
NEITHER
NICE
NO
NONE
NOR
NOT
NOTHING
NOW
OBJECT
OF
OFF
ON
ONE
ONLY
ONTO
OR
OUT
OVER
PICK
PLEASE
PUT
POINTER
PYRAMID
RED
RELEASE
RIGHT
ROUND
SAW
SEE
SET
SHAPE
SHE
SHORT
SINCE
SIT
SIZE
SMALL
SOME
SOMETHING
SPHERE
SQUARE
STACK
SUPPORT
TABLE
TAKE
TALL
TELL
THAN
THANK
THAT
THE
THEIR
THEM
THEN
THERE
THEY
THICK
THIN
THING
THIS
THREE
TIME
TO
TOGETHER
TOLD
TOP
TOUCH
TOY
TWO
UNDER
UNDERNEATH
UP
US
WANT
WAS
WERE
WHAT
WHEN
WHERE
WHICH
WHILE
WHITE
WHO
WHOM
WHOSE
WHY
WIDE
WILL
WITH
WOULD
YOU
YOUR

TABLE 2. The 200-word Lexicon Adapted from Winograd (1971)

1. Put the pyramid on the blue block.
2. Pick up a block in the box.
3. Pick up the block and the sphere.
4. The big block is on the table.
5. If the cube is not blue, pick it up.
6. Put it down if it is a cube.
7. Stack up all the objects in a box.
8. Put it on a block or in the box.
9. Why did you pick the blue block up?
10. Are there two blocks in a green one?

TABLE 3. Ten Sentences Constructed from the Winograd Lexicon. Words that are Underlined were Missed by one Experimenter. A Double Underline Indicates the Three Words that Were Missed by Both Experimenters

KNS Guess	Correct	DHK Guess	Correct
		a	the
a	the	a	the
a	the	a	the
the	a	a	the
every	a green	a	the
black	blue	in	and
all below*	not blue	[no guess]*	not
both of the cubes	all of the objects	than...does*	down if it is
		green*	or in
		he	you
		box	block
		the	blue
		put the	are there

*Word(s) do not make well-formed sentence so presence of an error was known.

TABLE 4. A Breakdown of the Errors Made by the Experimenters in the Sentence Recognition Task

3.2 Outline of Recognition Strategies

Our recognition strategies seemed to involve three steps. As a first step, while attempting to work in a left-to-right fashion, we would first identify certain clear and well-defined phonetic segments and features. These features are identified on the basis of spectra and spectral changes extending over only a brief interval, probably only a few tens of milliseconds. (The remaining features seem to be characterized by context-dependent acoustic attributes whose decoding was either impossible or required very complex reasoning involving acoustic data extending over a longer time interval, possibly one second or more.) The second step involved hypothesizing values for the remaining features from our theoretical knowledge of acoustic phonetics, by reference to the lexicon and through our intuition concerning syntactic and semantic constraints. In the final step, we determined whether the acoustic data were consistent with the hypothesized word sequence and feature values. The results of this step would either be a very satisfying discovery that all of the varied acoustic cues seemed to fit together or an uncomfortable inability to resolve some conflicting aspects of the data, in which case other alternatives were hypothesized.

3.2.1 An Example of the Recognition Process

In order to clarify this process, an example of one of the spectrograms is shown in Fig. 4. The utterance probably begins with a /p/ followed by a short vowel, a flap /d/, a short front vowel, and then a dental stop with a long closure duration. At this point, a quick lexical search revealed no multi-syllable word having a good match to the partial feature specification,

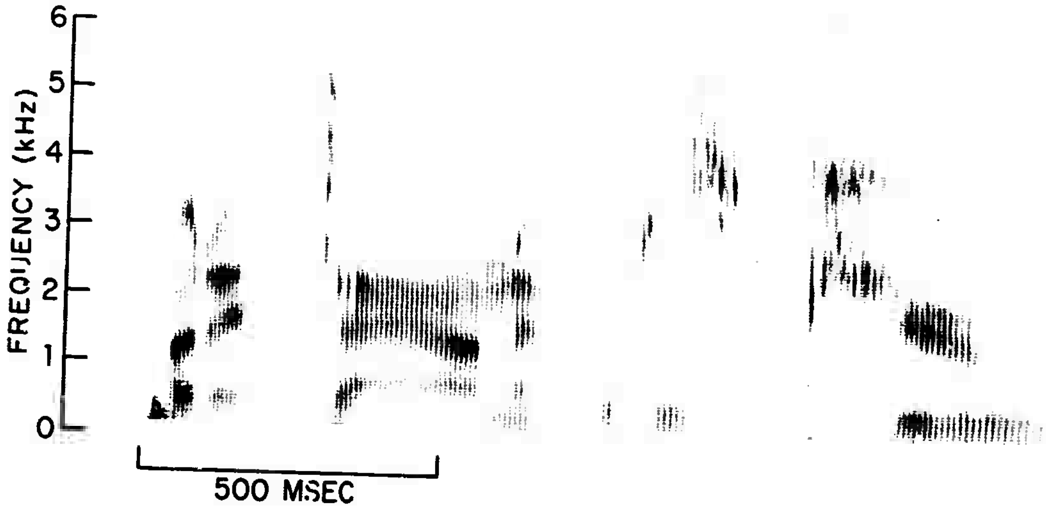


FIG.4 A BROADBAND SPECTROGRAM OF UNKNOWN SENTENCE NO. 6

although the word "pointer" was perhaps close enough that further analysis of the spectrogram was needed to reject it with confidence. Possible 2-word sequences were then considered and "put it" suggested itself as a good candidate. After consideration of the likely coarticulation pattern on the 2-word sequence reference to the spectrogram revealed that the vowel transitions and consonant cues "looked right" for this word pair. If no satisfactory lexical candidate had been found, we would have looked again at the data to see if perhaps the initial /p/ was a /b/ or an /f/, etc., and try the lexicon again, or we might have looked ahead to the next syllable containing a stressed vowel.

The next portion of the utterance probably begins with a /d/ (high frequency burst) followed by a long, low front vowel. The vowel is either diphthongized or prenasalized, and it is followed by a nasal, a short vowel and a stop. As with most stops or silent intervals, we must consider the possibility that a weak fricative such as /f,ð/ is present but not visible. Lexical searches of one- and two-word sequences revealed no likely candidates. "Handle" is rejected for several inconsistencies. In an expanded lexical search with relaxed matching criteria, the word "down" was found. Critical reasoning that included the possibility of a slight speaker accent on the vowel nucleus suggested that this was a satisfactory match, and the process continued.

The final syllable contains several good examples of how phonetic segments interact to obscure their respective identities if one uses simple decoding rules. The syllable starts with silence, followed by a plosive burst. The burst is probably a velar release, since it has a concentration of energy at about 2 KHz. It is followed by either a prolonged aspiration interval or a fricative such as /š/. The vowel nucleus begins as a high-front vowel which is diphthongized. There may or may not be a final consonant. This summary of the superficial aspects of the pattern leaves one with a very unsatisfactory feeling. The vowel transition is not typical of any stressed English glide or diphthong, and the aspiration (if it is aspiration) is much longer than one would expect in a plosive-vowel transition. With this preliminary analysis, it is difficult to go to the lexicon and find the correct word, i.e., "cube." However, if this word is considered as a possibility (as it might be if the syntactic or semantic context is taken into account) everything falls into place. Prolonged aspiration with some stimulus frication generation is typical of voiceless plosive plus /y/ coarticulations, and the prolonged transition of /y/ plus vowel that never seems to reach the /u/ target is also reasonable from articulatory considerations. A simple automatic procedure could easily take this sequence and form the syllable /kši/ because the acoustic data would probably match this sequence of segments reasonably well.

3.2.2 Acoustic Attributes used in Recognition Strategy

A partial list of acoustic attributes is given in Table 5, to indicate the types of spectrographic patterns that we tended to focus upon during the initial feature analysis. Formant-frequency changes also played a significant role in the analysis due to the fact that many of the acoustic-phonetic rules that we attempted to apply to the data were originally learned in terms of formant parameters. Formant frequencies are not always synonymous with spectral energy concentrations, particularly if

- PERIODIC SOUND
- PRESENCE OF NOISE
- RAPID SPECTRUM CHANGE
- RAPID ONSET OF ENERGY
- SILENCE
- SLOW SPECTRUM CHANGE
- LOCATIONS OF MAJOR SPECTRAL ENERGY CONCENTRATIONS
- CHANGE IN FUNDAMENTAL FREQUENCY

TABLE 5. A Partial List of Acoustic Properties used as Visual Cues

two formants are close together and a broad analyzing filter is used. Presumably, the rules could be restated in terms of major spectral energy concentrations in an automatic speech recognition application because formants are often difficult to determine automatically from the acoustic waveform.

Note that we have chosen to de-emphasize the idea of locating segment boundaries as an independent step in the analysis. Some types of boundaries fall naturally into the class of acoustic attributes mentioned above; other segment boundaries were not explicitly located in time as a part of the analysis.

3.2.3 Phonetic Features

We have talked about phonetic features of the type proposed by Chomsky and Halle (1968) or Jakobson, Fant and Halle (1963) during this paper. An example of a possible feature matrix for the lexical ‹ try "second" is shown in Table 6. The presence or absence of a feature is indicated by a + or a - in the appropriate place in the table. There are certain features which apply only if other features are present in a segment. For example, for a segment that is [+ labial] or [+ coronal] in English, the features high, low, back and rounded do not apply (or are predictable from the context). Or, the features anterior and lateral apply to [+ coronal] segments. The features in Table 6 are a subset of those proposed by Chomsky and Halle (1968), except for some minor changes.

An optimum set of features for automatic speech recognition purposes has yet to be developed. We have no reason to suppose, however, that it will be substantially different from the phonetic features that play such a powerful role in expressing the phonological constraints in language, as shown, for example,

	s	ɛ	k	ə	n	d
SYLLABIC	-	+	-	+	-	-
CONSONANTAL	+	-	+	-	+	+
STRESS		+		-		
CONTINUANT	+		-		-	-
SONORANT	-	+	-	+	+	-
NASAL	-		-		+	-
STRIDENT	+		-		-	-
VOICED	-	+	-	+	+	+
HIGH		-	+	-		
LOW		-		-		
BACK		-		-		
ROUNDED		-		-		
LABIAL	-		-		-	-
CORONAL	+		-		+	+
ANTERIOR	+				+	+
LATERAL	-				-	-
TENSE		-		-		

TABLE 6. A Feature Matrix for the Lexical Entry "Second." Phonetic Feature Values for Six Segments are Specified when Applicable. The Feature Values to be Expected in an Actual Realization of the Word Depend on the Sentence to be Spoken in Accordance with the Generative Rules of English Phonology

by Chomsky and Halle (1968). As we have indicated, some phonetic features may be rather directly related, through simple rules, to acoustic attributes of the type just described. Other features are more abstract in the sense that their acoustic correlates may be greatly influenced by the context.

The advantages of a feature representation over a traditional phonemic representation are significant. Features form a natural language for expressing partial information about a phonetic segment. A feature organization aids in performing sophisticated lexical searches by allowing questions such as "Give me all lexical entries containing a strident followed by a front vowel." All of the rule-governed transformations that take place when words combine to form an utterance, such as coarticulation, segment deletion, feature changes, durational changes, and word stress effects are described far more easily in terms of features than for example in terms of lists of phonemes.

4.0 CONCLUSIONS

Our experiments were very limited in time and materials, but certain conclusions can be drawn. From the first experiment, one is left with the impression that a significant error and omission rate is inevitable in a pure phonetic transcription task. A phonetic transcription of this type, if successfully implemented as a computer program, could be used to generate hypotheses about lexical items appearing in an unknown utterance, very much like our initial analysis in terms of relativly unencoded features aided in proposing potential word strings. However, we then found it necessary to go back to the primary acoustic data in order to tell whether a hypothesis should be accepted or rejected. With the high feature error and omission rates that are likely in an automated phonetic transcription procedure, it seems reasonable to believe that a similar type of hypothesis-verification process involving the primary acoustic data will be needed. A phonetic transcription which contains 10 to 15 percent errors and which leaves a number of features unspecified simply becomes too ambiguous to be decoded by higher-level programs unless very powerful syntactic, semantic and lexical constraints apply (Hanne and Shoup, 1965).

The results of the second experiment suggest that visual recognition of spectograms from a 200-word lexicon can be done with a fairly small error rate. But, is this an encouraging result for workers in the field of automatic speech recognition? We have the subjective feeling that it is not. The reason is the seeming complexity of the things we were doing in our heads in order to recognize a feature or word or phrase. It is not simply the enormous number of detailed facts that one must learn and know. The number of facts is incredibly large and not well documented, but this is not in principle an obstacle for the computer systems of today. What is staggering is the magnitude and complexity of the semantic and syntactic information that is available in the long-term memory of a human observer, and the complexity of our reasoning as we manipulate the facts at all levels to assess what is possible and what is not possible.

Is this reasoning power necessary in order to decode context-dependent features and to validate hypotheses that are generated? For an application such as the one described which involves a 200-word lexicon and a relatively open syntax, we think that it is indeed necesary.

4.1 Comparison with the State-of-the-Art

In order to put these thoughts into concrete terms, it is instructive to consider an example of a state-of-the-art speech recognition device. Vicens and Reddy have designed a system that controls a robot by recognizing sentences constructed from a 16-word lexicon with rigid syntactic constraints (Vicens, 1969; Reddy, 1967). It is probably the best (if not the only) device that has been built to date that deals with continuous speech. We shall not describe in detail the various steps that take place in their recognition strategy, but certain analogies can be drawn between their work and the recognition framework that we have described. The phonemic categories that they use are very similar to our so-called obvious phonetic feature distinctions. Context-dependent phonetic distinctions such as place-of-articulation for stops are not even attempted by Vicens and Reddy.

Instead, the matching of the input to the lexicon is done in terms of a course phonemic classification based on the raw acoustic data. A second point to note is that there is nothing analogous to our process of going back to the acoustic data to check the consistency of an hypothesis. That is, after initial phonetic categorizations are made, the acoustic data are not preserved in a "pre-categorical" store for future analysis. Sophisticated phonetic feature decoding rules and a hypothesis verification stage are not needed in the Vicens-Reddy application because the lexicon is small enough (16 words) and the syntax is very constraining. It is our feeling that this type of recognition system will have to be modified significantly in order to extend to larger vocabularies because the only current mechanism for dealing with errors in the input representation is to restrict the number of possibilities at any decision point to a small enough number that gross acoustic distinctions suffice most of the time.

4.2 Program of Future Research

This pilot study has only begun to describe the potential problems to be faced by a continuous speech recognition device. Continued work with spectrographic data appears to be a rich source for developing increased knowledge about the nature of these problems. Questions that remain concerning the visual recognition task include: (1) Do new speakers require recalibration of our decision criteria? (2) Will practice at this task tend to change our strategies and reveal short-cuts? (3) How does one begin to formalize our protocols and collect specific facts about English in computer-implementable form?

In order to investigate these questions, we plan to select a new lexicon based on a potentially useful continuous speech recognition application (Woods, 1971). The lexicon will be stored in the computer in terms of segments and features (see Table 6). A simple user-oriented language is being developed to make possible feature-based questions about lexical entries. This language will facilitate scans of the lexicon in future visual recognition experiments.

A collection of sentences covering the vocabulary will be recorded by several speakers and broadband spectrograms will be made. In working with this new material, we will verbalize our thought processes and save a list of questions asked about the lexicon. A subsequent protocol analysis will be performed in the hope of formalizing a recognition strategy and improving the form of the lexical representation.

4.3 The Problem of Machine Implementation

In some respects, a broadband spectrogram is not an optimum form of representation for visual recognition of speech. The limited dynamic range from blackest black to lightest grey in a spectrogram means that many of the weaker consonants are poorly represented in a spectrographic display. This need not be limitation for the representation in a computer as long as a good signal-to-noise ratio is preserved in the original acoustic data. However, the problem of automatic extraction of acoustic properties such as formant frequencies, fundamental frequency, rapid spectral changes, etc., remains as a serious obstacle to the machine implementation of any recognition strategy. It is also true that many of the detailed facts about the acoustic phonetics of English are not available in machine implementable form. However, a carefully selected program of reading in this area (Rothenberg, 1963; Fant, 1960; Lehiste, 1968) can lay the foundation for serious work on speech recognition systems.

In conclusion, it is suggested that every serious worker in the area of automatic speech recognition should undertake to read spectrograms in an organized way similar to the projects that we have described. It is an excellent way of learning a great deal about speech, and it is the only sure way to convince yourself of the complexities involved and of the necessity for approaching the problem with more sophisticated forms of analysis.

5.0 REFERENCES

Chomsky, N. and Halle, M. *Sound Pattern of English*. New York: Harper and Row, 1968.

Fant, C. G. M. *The Acoustic Theory of Speech Production*. Mouton and Co., 's-Gravenhage, 1960.

Gold, B. *Word Recognition Computer Program*. Technical Report 452, Research Laboratory of Electronics. Cambridge, Mass.: M.I.T., 1966.

Goldberg, A. J. *Visual Speech Displays for the Severely Hard of Hearing*, Ph.D. Thesis, M.I.T., Cambridge, Mass., 1970.

Hanne, J. and Shoup, J. E. *The Problem of Determing the Words of a Sentence from a Phonetic Transcription*. Unpublished study, Speech Communications Research Laboratory, Santa Barbara, Calif., 1965.

Hyde, S. R. *Automatic Speech Recognition Literature Survey and Discussion*. Research Department Report No. 45, Joint Speech Research Unit, London, England, 1968.

Jakobson, R., Fant, G., and Halle, M. *Preliminaries to Speech Analysis*. Cambridge, Mass.: M.I.T. Press, 1963.

Kersta, L. G. Voiceprint identification. *Nature*, 196, 1253-1257, 1962.

Koenig, W., Dunn, H. K. and Lacy, L. Y. The sound spectograph. J. Acoust. Soc. Am., 17, 19-49, 1946.

Lehiste, I. Selected Readings In Acoustic Phonetics. Cambridge Mass.: M.I.T. Press, 1968.

Liberman, A. M., Cooper, F. S., Shankwaiter, D. O. and Studdert-Kennedy, M. Why are speech spectograms hard to read? Am. Annals of the Deaf, 113, 127-133, 1968.

Lindgren, N. Automatic speech recognition. IEEE Spectrum, 2, 114-136, 1965.

Martin, T. G., Nelson, A., Zadell, H. and Cox, R. Continuous Speech Recognition by Feature Abstraction. DDC No. AFAL-78-66-169, Camden, N.J.: Radio Corporation of America, 1966.

Newell, A. (ed) Final Report of a Study Group on Speech Understanding Systems, prepared for the Advanced Research Topics Agency of the Department of Defense, in press.

Potter, R. K., Kopp, G. A. and Green, H. C., Visible Speech. D. Van Nostrand Co., Inc., 1947, and New York: Dover Publications, Inc., 1966.

Presti, A. J. High speed sound spectrograph. J. Acoust. Soc. Am., 40, 628-634, 1966.

Reddy, D. R. Computer recognition of connected speech. J. Acoust. Soc. Am., 42, 329-347, 1967.

Rothenberg, M. *Programmed Learning Problem Set to Teach the Interpretation of a Class of Speech Spectrograms*. Ann Arbor, Michigan: Ann Arbor Publishers, 1963.

Sakai, T. and Doshita, S. The Phonetic Typewriter, *Information Processing*, 1962, Proc. IFIP Congress, Munich, August-September, 1962.

Stevens, K. N. *Summer Program in Speech Communication*. Course 6.69S, Massachusetts Institute of Technology, Cambridge, Massachusetts, June 23-July 3, 1969.

Tappert, C. C., Dixon, N. R., Beetle, D. H. and Chapman, W. D. *The Use of Dynamic Segments in the Automatic Recognition of Continuous Speech*, Technical Report RADC-TR-70-22, IBM, Systems Development Division, Research Triangle Park, N.C., 1970.

Tosi, O., Oyer, H., Pedrey, C., Lashbrook, B. and Nicol, J. An experiment on voice identification by visual inspection of spectograms. *J. Acoust. Soc. Am.*, 49, 138(A), 1971.

Vicens, P. *Aspects of Speech Recognition by Computer*, Ph.D. Thesis, Technical Report Number CS-127, Computer Sciences Department, Stanford University, Palo Alto, Calif., 1969.

Winograd, T. *Procedures as a Representation of Knowledge in a Computer Program for Understanding Natural Language*. Artificial Intelligence Memo, Project Mac, Massachusetts Institute of Technology, Cambridge, Mass., 1971.

Woods, W. The Lunar Sciences Natural Language Information System, Internal Report under Contract NAS9-1115, Bolt Beranek and Newman, Inc., Cambridge, Massachusetts, 1971.